

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-283184
(43)Date of publication of application : 12.10.2001

(51)Int.Cl.

G06N 3/00

G06F 9/44

G06F 17/30

(21)Application number : 2000-091863

(71)Applicant : MATSUSHITA ELECTRIC IND CO LTD

(22)Date of filing : 29.03.2000

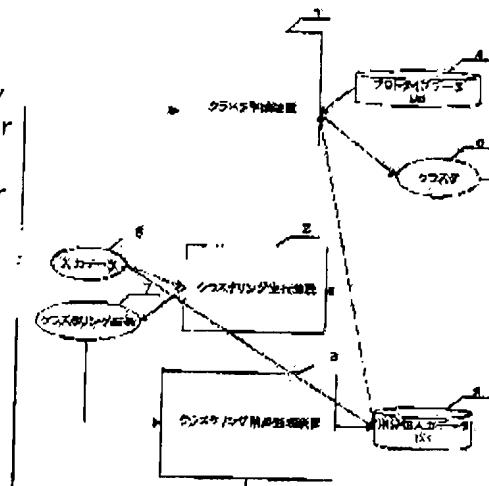
(72)Inventor : NAKAMITSU HIROAKI

(54) CLUSTERING DEVICE

(57)Abstract

PROBLEM TO BE SOLVED: To provide a clustering device capable of coping with the dynamic change of data in clustering in a simple constitution and procedure.

SOLUTION: This clustering device for classifying input data by using a cluster is provided with a cluster preparing device 1 for preparing a cluster, a clustering performing device 2 for performing the clustering of the input data by using the cluster prepared by the cluster preparing device, a clustering result monitoring device 3 for monitoring the clustering result of the clustering performing device, and for identifying the erroneously classified input data, and a storage means 8 for storing the erroneously classified input data. When the fixed number of data or more are stored in the storage means, a new cluster is prepared by the cluster preparing device based on the data. Thus, it is possible to correct the cluster corresponding to the dynamic change of the input data, and to reduce the erroneous classification.



* NOTICES *

JPO and INPI are not responsible for any damages caused by the use of this translation.

1. This document has been translated by computer. So the translation may not reflect the original precisely.
2. **** shows the word which can not be translated.
3. In the drawings, any words are not translated.

CLAIMS

[Claim(s)]

[Claim 1] A clustering device characterized by said cluster preparation device creating a new cluster based on said data when it has the following and data more than fixed numbers is stored in said accumulation means.

A cluster preparation device which is a clustering device which classifies input data using a cluster, and creates said cluster.

A clustering performing device which performs clustering of input data using a cluster created by said cluster preparation device.

A clustering result monitoring instrument which identifies input data by which supervised a clustering result of said clustering performing device, and false sorts were carried out.

An accumulation means which accumulates said input data by which false sorts were carried out.

[Claim 2] The clustering device according to claim 1 adding said cluster preparation device to a cluster which used said data, created a new cluster automatically, and was already created when data more than fixed numbers is stored in said accumulation means.

[Claim 3] The clustering device according to claim 1, wherein data of an error of clustering is contained in said clustering result.

[Claim 4] The clustering device comprising according to claim 1:

A self-organization map creating means which said cluster preparation device considers prototype data as an input, and generates a self-organization map.

A cluster formation means to classify said generated self-organization map and to form a cluster.

[Claim 5] The clustering device according to claim 4 which is provided with the following and characterized by adding a cluster formation means of said cluster preparation device to a cluster which classified the self-organization map, created a cluster, and was already created when said self-organization map correcting means generates a self-organization map.

A clustering result monitor means in which said clustering result monitoring instrument supervises a clustering result.

A self-organization map correcting means which generates a self-organization map by considering said data as an input when data more than fixed numbers is stored in said accumulation means.

[Translation done.]

* NOTICES *

JPO and INPI are not responsible for any damages caused by the use of this translation.

1. This document has been translated by computer. So the translation may not reflect the original precisely.
2. **** shows the word which can not be translated.
3. In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]

[0001]

[Field of the Invention] This invention enables it to correspond to the dynamic change of input data appropriately especially about the clustering device which classifies much data into a class from the similarity.

[0002]

[Description of the Prior Art] Conventionally, various things are proposed as the clustering technique. The example of the most common clustering device is shown in drawing 6.

[0003] 100 is shown and here the prototype data constellation for study 102 and 103, The bottom indicates the cluster A and the cluster B to be initial clusters for each data of a prototype data constellation wholly, 104 shows the distance of the cluster A102 and the cluster B103, and 105 shows the cluster C which unified the cluster A102 and the cluster B103. 200 shows the cluster result created from prototype data, and 201 and 202 show the cluster Y created eventually and the cluster Z. 300 shows the clustering device which used the cluster, and the cluster Y with 301 completely of the same type as 201 and the cluster Z with 302 completely of the same type as 202 are shown, and 303. The input X which is an object of clustering is shown and 304 shows the point of the input X303 of Jo Sorama in whom a cluster exists.

[0004] In this device, a cluster required for a clustering device is created first. This is called for by the following work.

[0005] Supposing it looks for a cluster with the nearest distance and the cluster A102 and the cluster B103 are chosen from the prototype data constellation 100 for study as a result, these two will be unified, it will be considered as the cluster C105, and the clusters A and B will delete. At this time, the cluster C105 has a value of the cluster A102 and the cluster B103 both. Next, a series of work of looking for a cluster with the nearest distance and unifying them from the prototype data constellation 100 similarly is repeated. Work is ended, when the total cluster number was set to 1 at this time, or when the distance of clusters with the nearest distance is larger than a certain constant value.

[0006] By this work of a series of, the cluster result 200 created from prototype data is searched for, and the cluster integrated eventually turns into the cluster Y201 and the cluster Z202.

[0007] The clustering device 300 using a cluster performs actual clustering using these clusters integrated eventually. When the input X303 is inputted into the clustering device 300 using this cluster and the input X303 is included in the cluster Y301, the input X303 brings the result of having been clustered by the cluster Y301.

[0008] clustering --- a self-organization map (SOM:Self-Organization Map --- in detail) T. Kohonen, It is indicated to "Self-Organization and Associative Memory", Third Edition, Springer-Verlag, Berlin, and 1989. The neural network called. The technique to be used is also known (JP,7-234853,A). Prototype data is inputted into SOM, the neurone which forms SOM is learned and the learned neurone is classified into a cluster according to this method. If input data is given to SOM after a cluster is formed, neurone with the value near the input will be determined, and input data will be clustered.

[0009]

[Problem(s) to be Solved by the Invention] However, in the above clustering techniques, since the

cluster is formed using prototype data, the cluster which inclined only toward prototype data is formed. Therefore, when live data are actually clustered using these clusters, there is a problem referred to as being unable to respond to a dynamic change of input data.

[0010]That is, when the data which should belong to a new class arises with the passage of time, in the conventional method, correspondence will be impossible at all and false sorts will be carried out to one of clusters.

[0011]In order to prevent these false sorts, it is necessary to recluster using all the data, and is forced a big work burden in the conventional method also including prototype data. When data is newly added, the method of correcting a cluster is indicated by JP,5-205058,A, but, It must be known to have added new data, and this needs to tell making correction of the cluster by addition of data from the exterior, and the data to add cannot be collected automatically or it cannot correct a cluster automatically.

[0012]This invention solves such a conventional problem, they are easy composition and a procedure and an object of this invention is to provide the clustering device which can respond to the dynamic change of the data in clustering.

[0013]

[Means for Solving the Problem]Then, in a clustering device which classifies input data according to this invention using a cluster, A cluster preparation device which creates a cluster, and a clustering performing device which performs clustering of input data using a cluster created by a cluster preparation device, A clustering result monitoring instrument which identifies input data by which supervised a clustering result of a clustering performing device and false sorts were carried out, When an accumulation means which accumulates input data by which false sorts were carried out is established and data more than fixed numbers is stored in an accumulation means, it constitutes based on this data so that a cluster preparation device may create a new cluster.

[0014]Therefore, a cluster can be corrected corresponding to a dynamic change of input data, and false sorts can be stopped.

[0015]

[Embodiment of the Invention]Hereafter, an embodiment of the invention is described using a drawing. This invention is not limited to these embodiments at all, and can be carried out in the mode which becomes various in the range which does not deviate from the gist.

[0016](A 1st embodiment) The clustering device of a 1st embodiment is provided with the following.
Prototype data DB4 which manages prototype data as shown in drawing 1.

The cluster preparation device 1 which creates the cluster 5 using prototype data.

The clustering performing device 2 which clusters the input data 6 using the created cluster 5.

False-sorts input data DB8 which manages the data judged to be false sorts with the clustering result monitoring instrument 3 which supervises the clustering result 7 of the clustering performing device 2, and the clustering result monitoring instrument 3.

[0017]In this device, the cluster preparation device 1 generates the cluster 5 using prototype data DB4. The clustering performing device 2 clusters the inputted input data 6 using the generated cluster 5, and outputs the clustering result 7. The clustering result monitoring instrument 3 supervises the outputted clustering result 7. When it judges that the error included in the clustering result 7 of the input data 6 is a value beyond a certain constant value, and is false sorts clearly, the input data 6 is added to false-sorts input data DB8, and the number of data collected on false-sorts input data DB8 is counted. When fixed numbers with the data in this false-sorts input data DB8 are exceeded, this false-sorts input data DB8 is used for the cluster preparation device 1, and it directs to create a cluster.

[0018]Operation of each device is explained in more detail. First, the cluster preparation device 1 operates, when the cluster 5 is not created, and when creation of a cluster is directed from the clustering result monitoring instrument 3.

[0019]When the cluster 5 is not created, it considers that each data of the prototype data constellation in the prototype data DB4 is an initial cluster, and a cluster with the nearest distance is looked for out of it. This distance is found by the formula 1 of drawing 5. Two clusters called for at this time are unified, and it is considered as a new cluster. The cluster which deleted the cluster integrated and was newly made has all the values of the cluster deleted by integration. A series of work of looking for a

cluster with the nearest distance and unifying them from the prototype data DB in a similar manner again is repeated. Work is ended, when the total cluster number was set to 1 at this time, or when the distance of clusters with the nearest distance is larger than a certain constant value.

[0020]According to this work of a series of, the cluster 5 created from the prototype data 4 is created.

[0021]Next, when directions of cluster creation are received from the clustering result monitoring instrument 3, using false-sorts input data DB8, it is the same operation as creating the cluster 5, and a cluster is created. At this time, by the created cluster, the number of the values contained in a cluster makes the thing more than fixed a new cluster, and it adds to the cluster 5. Finally, the false-sorts input data DB is cleared.

[0022]Next, operation of the clustering performing device 2 is explained. The inputted input data 6 and the nearest cluster of distance are chosen using the cluster 5 created by the cluster preparation device 1. It asks for calculation of this distance by the formula 1 of drawing 5. At this time, the selected cluster and the calculated distance showing an error are outputted as the clustering result 7.

[0023]The error included in the clustering result 7 to which the clustering result monitoring instrument 3 was outputted, Namely, when the calculated distance is a value beyond a certain constant value, add the input data 6 to false-sorts input data DB8, and the number is counted. When fixed numbers with the data in this false-sorts input data DB8 are exceeded, this false-sorts input data DB8 is used for the cluster preparation device 1, and it directs to create a cluster.

[0024]As mentioned above, in the clustering device of this embodiment, also during operation, automatic creation of a cluster is possible and a cluster can be automatically created corresponding to a dynamic change of input data. Therefore, generating of the false sorts resulting from a dynamic change of input data is suppressed promptly. In this device, since re-creation of a cluster is performed only using the data by which automatic collection was carried out as false-sorts data in process of clustering of live data, correction of a cluster can be made at little burden.

[0025](A 2nd embodiment) The clustering device of a 2nd embodiment creates a cluster using a self-organization map (henceforth SOM).

[0026]This device is provided with the following.

The data input means 11 as which prototype data DB4, cluster preparation device 1, clustering performing device 2, clustering result monitoring instrument 3, and false-sorts input data DB8 is comprised, and the cluster preparation device 1 inputs prototype data like a 1st embodiment as shown in drawing 2.

The SOM preparing means 12 which creates SOM9.

The clustering result monitor means 31 which is equipped with the cluster creating means 13 which generates a cluster using SOM9 and in which the clustering result monitoring instrument 3 supervises the clustering result 7 of the clustering performing device 2.

The SOM correcting means 32 which creates SOM10 using the data of false-sorts input data DB8.

[0027]In this device, the data input means 11 of the cluster preparation device 1 inputs data from prototype data DB4, the SOM preparing means 12 creates SOM9 using this data, and the cluster creating means 13 generates the cluster 5 using SOM9. The clustering performing device 2 clusters the input data 6 inputted using the generated cluster 5, and outputs the clustering result 7. When it judges that the error included in the outputted clustering result 7 is a value beyond a certain constant value, and is false sorts clearly, the clustering result monitor means 31 of the clustering result monitoring instrument 3 adds the input data 6 to false-sorts input data DB8, and counts the number.

[0028]When fixed numbers with the data in the false-sorts input data DB8 are exceeded, the SOM correcting means 32 creates SOM10 [new as an input] for the data of false-sorts input data DB8, and directs the cluster creation which used SOM10 for the cluster preparing means 13. In response, the cluster preparing means 13 is added to the cluster 5 which creates a cluster using SOM10 and has already been created.

[0029]Next, it explains in more detail about operation of each part. First, operation of the SOM preparing means 12 is explained.

[0030]As SOM is shown in drawing 4, it is formed from the neurone 402 arranged on two dimensions, and each neurone has a vector of the same dimension as the input called the reference vector 403.

[0031]The SOM preparing means 12 creates SOM in the procedure shown in the flow chart of drawing

3.

Step A1: Set the learning frequency T to 0, create the neurone arranged on two dimensions like step A2:drawing 4, and give the reference vector of the same dimension as an input by random numbers to each neurone.

[0032]Step A3: It is random from prototype data DB4, and the data input means 11 takes out one data.

[0033]Step A4: Determine the neurone C with the reference vector which fills the formula (2) of drawing 5 to this data.

[0034]Step A5: Update the reference vector of the neurone located near the neurone C according to the formula (3) of drawing 5.

[0035]Step A6: When the number of times which the learning frequency T specified is reached, do the end of step A8:of.

[0036]In Step A6, when the learning frequency T has not reached the number of times of regulation, one value of the step A7:learning frequency T is increased, and it returns to Step A2.

[0037]Next, the cluster creating means 13 operates, when the cluster 5 is not created, and when directions of creation of a cluster are received from the SOM correcting means 32.

[0038]First, when the cluster 5 is not created, the cluster 5 is created using SOM9. To each neurone of SOM9, neurone with the reference vector which fills the formula (4) of drawing 5 is chosen, and it is considered that the selected neurone is an initial cluster. A cluster with the nearest distance is looked for out of it. This distance is found by the formula (1) of drawing 5. Two clusters called for at this time are unified, and it is considered as a new cluster. The cluster which deleted the cluster integrated and was newly made has all the values of the cluster deleted by integration. The cluster whose distance is again the nearest is looked for similarly, and a series of work of unifying them is repeated. Work is ended, when the total cluster number was set to 1 at this time, or when the distance of clusters with the nearest distance is larger than a certain constant value.

[0039]When transfer of creation of a cluster is received from the SOM correcting means 32, similarly, a cluster is created using SOM10 and it adds to the cluster 5.

[0040]Like a 1st embodiment, the clustering performing device 2 clusters the input data 6 using the cluster 5, and outputs the clustering result 7. When it judges that the error included in the outputted clustering result 7 is a value beyond a certain constant value, and is false sorts clearly, the clustering result monitor means 31 adds the input data 6 to false-sorts input data DB8, and counts the number.

[0041]When fixed numbers with the data in this false-sorts input data DB8 are exceeded, the SOM correcting means 32 creates small SOM10 with a size of a map equal to the number of the length of SOM9, or horizontal neurone by considering the data of false-sorts input data DB8 as an input according to the flow chart of drawing 3. And false-sorts input data DB8 is cleared and creation of a cluster is directed to the cluster preparing means 13. The cluster preparing means 13 creates a cluster using SOM10, and adds it to the created cluster 5 so that it may mention above.

[0042]As mentioned above, since it is clustering in the clustering device of this embodiment using SOM, since very small SOM is used when the existing SOM can be applied as it is and a cluster is still more newly created, processing speed is also high, and that practical effect is large. Since that by which automatic collection was carried out as false-sorts data in process of clustering of live data is used for creation of this new cluster, it can respond to a dynamic change of input data by creation of this new cluster.

[0043]

[Effect of the Invention]The clustering device of this invention can create a new cluster promptly corresponding to a dynamic change of input data, and can suppress generating of the false sorts resulting from a dynamic change of input data so that clearly from the above explanation.

[0044]Since creation of this new cluster is performed only using the data by which automatic collection was carried out as false-sorts data when clustering is performed, there are few those creation burdens and they end.

[0045]In a device with a means to create a cluster directly from the data by which false sorts were carried out, the advantageous effect that it is possible to create a cluster automatically also during device operation, and it can respond to a dynamic change of input data quickly is acquired.

[0046]In the device clustered using SOM, since very small SOM is used when the existing SOM can be applied as it is and a cluster is newly created, the effective effect that processing speed is also high is

acquired.

[0047]By this, input data can apply this invention to the device which clusters what changes in time, and an effect can be demonstrated. For example, it is very effective when it uses for the clustering device of the learning system which classifies a student by using as input data the learning result of the student who changes in time, the clustering device which investigates the palatability of the televiewer who accesses the homepage of the Internet, etc.

[Translation done.]

* NOTICES *

JPO and INPIT are not responsible for any damages caused by the use of this translation.

1. This document has been translated by computer. So the translation may not reflect the original precisely.
2. **** shows the word which can not be translated.
3. In the drawings, any words are not translated.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]

[Drawing 1] The block diagram showing the composition of the clustering device in a 1st embodiment of this invention,

[Drawing 2] The block diagram showing the composition of the clustering device in a 2nd embodiment of this invention,

[Drawing 3] The flow chart which shows the procedure of SOM creation in a 2nd embodiment,

[Drawing 4] The figure showing SOM visually,

[Drawing 5] The figure showing expression,

[Drawing 6] It is a figure showing an example of the conventional clustering device.

[Description of Notations]

- 1 Cluster preparation device
- 2 Clustering performing device
- 3 Clustering result monitoring instrument
- 4 Prototype data DB
- 5 Cluster
- 6 Input data
- 7 Clustering result
- 8 False-sorts input data DB
- 9, 10 SOM
- 11 Data input means
- 12 SOM preparing means
- 13 Cluster creating means
- 31 Clustering result monitor means
- 32 SOM correcting means
- 100 Prototype data constellation
- 102 Cluster A
- 103 Cluster B
- 104 Distance
- 105 Cluster C
- 200 The cluster result created from prototype data
- 201 Cluster Y
- 202 Cluster Z
- 300 A clustering device using a cluster
- 301 Cluster Y
- 302 Cluster Z
- 303 Input X
- 304 The point of the input X
- 401 SOM
- 402 Neurone
- 403 Reference vector

[Translation done.]

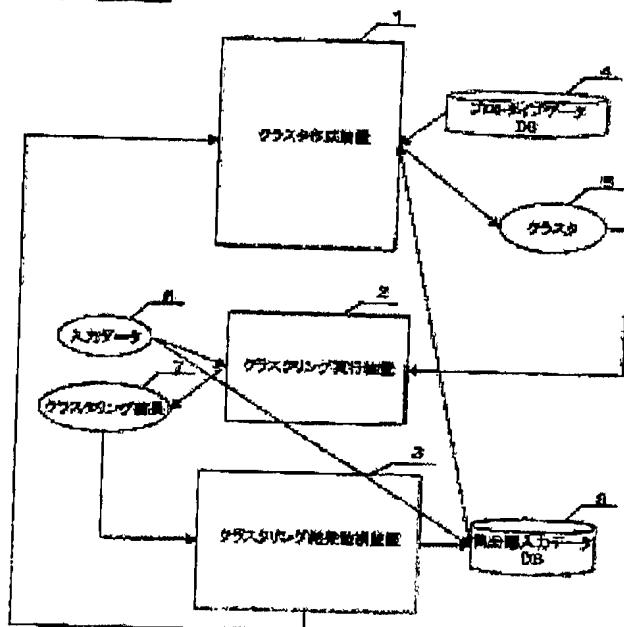
* NOTICES *

JPO and INPI are not responsible for any damages caused by the use of this translation.

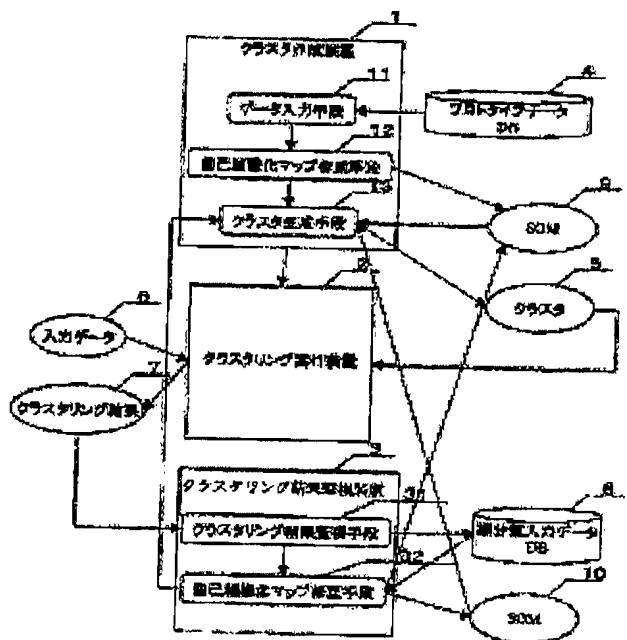
1. This document has been translated by computer. So the translation may not reflect the original precisely.
2. **** shows the word which can not be translated.
3. In the drawings, any words are not translated.

DRAWINGS

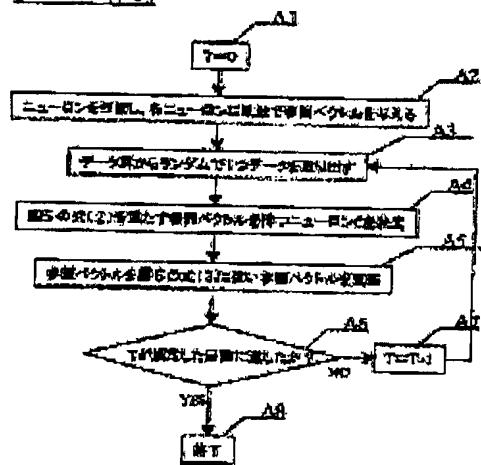
[Drawing 1]



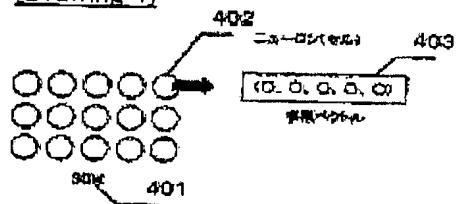
[Drawing 2]



[Drawing 3]



[Drawing 4]



[Drawing 5]

N次元のデータXYの距離D $D = \sqrt{\sum_{i=1}^N (x_i - y_i)^2}$ (m)

$C = \text{arg} \max_{\alpha} \prod_{j=1}^N h_{\alpha}(t_j) \exp(-\alpha \|m_j - m_{\alpha}\|)$

↓ニューロン番号
↓入力データ
↓ニューロンの参照ベクトル
↓式(1)によって求められる距離

$$m_{\alpha} = \begin{cases} m_1 + h_{\alpha}(t)(m_2(t) - m_1(t)) & t \in N_1 \\ m_2 & t \notin N_1 \end{cases} \quad (3)$$

$$h_{\alpha} = \alpha \exp \left(-\frac{\|m_{\alpha} - m_j\|^2}{2\sigma^2(t)} \right)$$

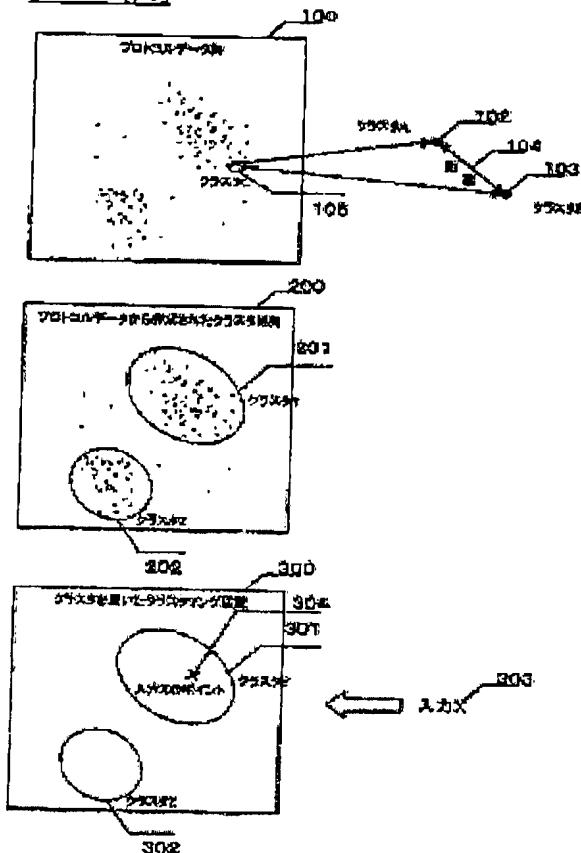
N_1 : ニューロンCの近傍に位置するニューロン
↑: ニューロン番号
α: 学習係数(0 < α < 1(時間とともに減少))
○: 近傍幅(時間とともに減少)
↓: 式(1)によって求められる距離
↓: 2次元マップ上のニューロンCの位置ベクトル
↓: 2次元マップ上のニューロンの位置ベクトル

$$d_j < \alpha \exp d_j \quad \text{if } j \in N_1$$

$$d_j = \frac{1}{\|m_j - m_{\alpha}\|} \quad \text{if } j \notin N_1$$

N_p : ニューロンCの4近傍中に存在するニューロンの集合
m_j: ニューロンjの参照ベクトル
m_p: ニューロンpの参照ベクトル
↓: 式(1)によって求められる距離

[Drawing 6]



[Translation done.]

(19)日本国特許庁 (JP)

(12) 公開特許公報 (A)

(11)特許出願公開番号
特開2001-283184
(P2001-283184A)

(43)公開日 平成13年10月12日 (2001.10.12)

(51) Int.Cl. ⁷	識別記号	F I	マーク* (参考)
G 06 N 3/00	5 6 0	G 06 N 3/00	5 6 0 A 5 B 0 7 5
G 06 F 9/44	5 8 0	G 06 F 9/44	5 8 0 A
17/30	2 1 0	17/30	2 1 0 D

審査請求 未請求 請求項の数 5 OL (全 7 頁)

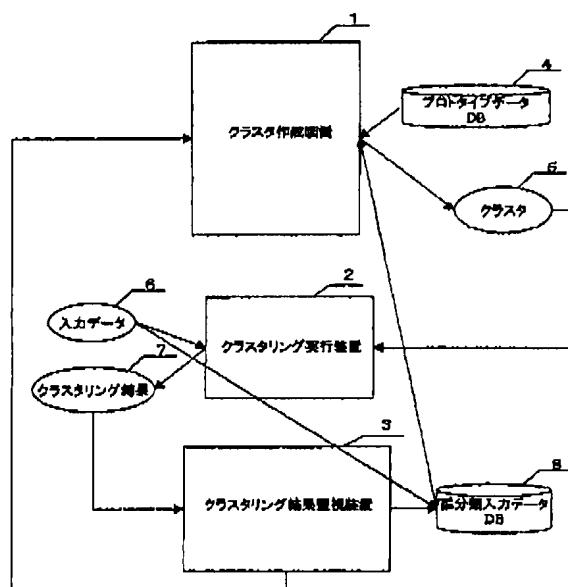
(21)出願番号	特願2000-91863(P2000-91863)	(71)出願人	000005821 松下電器産業株式会社 大阪府門真市大字門真1006番地
(22)出願日	平成12年3月29日 (2000.3.29)	(72)発明者	仲光 廣晃 大阪府門真市大字門真1006番地 松下電器 産業株式会社内
		(74)代理人	100099254 弁理士 役 昌明 (外3名) Pターム(参考) 5B075 NR12

(54)【発明の名称】 クラスタリング装置

(57)【要約】

【課題】 簡単な構成と手順で、クラスタリングにおけるデータの動的変化に対応することができるクラスタリング装置を提供する。

【解決手段】 入力データを、クラスタ用いて分類するクラスタリング装置において、クラスタを作成するクラスタ作成装置1と、クラスタ作成装置により作成されたクラスタを用いて、入力データのクラスタリングを実行するクラスタリング実行装置2と、クラスタリング実行装置のクラスタリング結果を監視して誤分類された入力データを識別するクラスタリング結果監視装置3と、誤分類された入力データを蓄積する蓄積手段8とを設け、蓄積手段に一定数以上のデータが蓄積された場合に、このデータを基に、クラスタ作成装置が新たなクラスタを作成するように構成している。入力データの動的変化に対応してクラスタを修正し、誤分類を抑えることができる。



【特許請求の範囲】

【請求項1】 入力データを、クラスタを用いて分類するクラスタリング装置であって、
前記クラスタを作成するクラスタ作成装置と、
前記クラスタ作成装置により作成されたクラスタを用いて、入力データのクラスタリングを実行するクラスタリング実行装置と、
前記クラスタリング実行装置のクラスタリング結果を監視して誤分類された入力データを識別するクラスタリング結果監視装置と、
誤分類された前記入力データを蓄積する蓄積手段とを備え、
前記蓄積手段に一定数以上のデータが蓄積された場合に、前記データを基に、前記クラスタ作成装置が新たなクラスタを作成することを特徴とするクラスタリング装置。

【請求項2】 前記クラスタ作成装置は、前記蓄積手段に一定数以上のデータが蓄積された場合に、前記データを用いて新たなクラスタを自動的に作成し、既に作成したクラスタに追加することを特徴とする請求項1に記載のクラスタリング装置。

【請求項3】 前記クラスタリング結果に、クラスタリングの誤差のデータが含まれることを特徴とする請求項1に記載のクラスタリング装置。

【請求項4】 前記クラスタ作成装置は、プロトタイプデータを入力として自己組織化マップを生成する自己組織化マップ生成手段と、生成された前記自己組織化マップを区分しクラスタを形成するクラスタ形成手段とを備えることを特徴とする請求項1に記載のクラスタリング装置。

【請求項5】 前記クラスタリング結果監視装置は、クラスタリング結果を監視するクラスタリング結果監視手段と、前記蓄積手段に一定数以上のデータが蓄積された場合に、前記データを入力として自己組織化マップを生成する自己組織化マップ修正手段とを備え、前記クラスタ作成装置のクラスタ形成手段は、前記自己組織化マップ修正手段が自己組織化マップを生成した場合、その自己組織化マップを区分してクラスタを作成し、既に作成したクラスタに追加することを特徴とする請求項4に記載のクラスタリング装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は、多数のデータをその類似性からクラスに分類するクラスタリング装置に関し、特に、入力データの動的変化に適切に対応できるようにしたものである。

【0002】

【従来の技術】 従来、クラスタリング手法として、さまざまなもののが提案されている。図6には、最も一般的なクラスタリング装置の例を示している。

【0003】 ここで、100は学習のためのプロトタイプデータ群を示し、102、103は、プロトタイプデータ群の個々のデータを初期クラスタとみなしたクラスタA、クラスタBを示し、104は、クラスタA 102とクラスタB 103との距離を示し、105はクラスタ A 102とクラスタ B 103とを統合したクラスタ Cを示す。200は、プロトタイプデータから作成されたクラスタ結果を示し、201、202は最終的に作成されたクラスタ Y、クラスタ Zを示す。300は、クラスタを用いたクラスタリング装置を示し、301は201とまったく同型のクラスタ Y、302は202とまったく同型のクラスタ Zを示し、303は、クラスタリングの対象である入力 X を示し、304はクラスタが存在する空間上の入力 X 303のポイントを示す。

【0004】 この装置では、まず、クラスタリング装置に必要なクラスタを作成する。これは以下の作業により求められる。

【0005】 学習のためのプロトタイプデータ群100から、最も距離の近いクラスタを探し、その結果、クラスタ A 102とクラスタ B 103とが選ばれたとすると、この2つを統合してクラスタ C 105とし、クラスタ A、Bは削除する。この時クラスタ C 105は、クラスタ A 102とクラスタ B 103との値を両方ともを持つ。次に、また同様にプロトタイプデータ群100から、最も距離の近いクラスタを探し、それらを統合する、という一連の作業を繰り返す。この時、全クラスタ数が1になった場合や、最も距離の近いクラスタ同士の距離が、ある一定値より大きかった場合は、作業を終了する。

【0006】 この一連の作業により、プロトタイプデータから作成されたクラスタ結果200が求められ、最終的に統合されたクラスタがクラスタ Y 201、クラスタ Z 202となる。

【0007】 これら最終的に統合されたクラスタを用い、実際のクラスタリングを行うのがクラスタを用いたクラスタリング装置300である。このクラスタを用いたクラスタリング装置300に入力 X 303が入力された時、入力 X 303がクラスタ Y 301内に含まれる時、入力 X 303は、クラスタ Y 301にクラスタリングされたという結果となる。

【0008】 また、クラスタリングに自己組織化マップ(SOM: Self-Organization Map、詳しくは、T. Kohonen, "Self-Organization and Associative Memory", Third Edition, Springer-Verlag, Berlin, 1989に記載されている。)と呼ばれるニューラルネットワークを用いる手法も知られている(特開平7-234853号)。この方法では、プロトタイプデータをSOMに入力して、SOMを形成するニューロンを学習し、学習したニューロンをクラスタに分類する。クラスタが形成された後、SOMに入力データを与えると、その入力に近い値を持つニューロンが決定され、入力データがクラスタリングされる。

【0009】

【発明が解決しようとする課題】しかし、前述のようなクラスタリング手法では、プロトタイプデータを用いてクラスタを形成しているので、プロトタイプデータにのみ偏ったクラスタが形成される。そのため、実際にこれらのクラスタを用いて実データのクラスタリングを行った時、入力データの動的な変化に対応できない、と云う問題点がある。

【0010】つまり、新たなクラスに属すべきデータが、時間の経過とともに生じた場合などに、従来の方法では、全く対応ができず、いずれかのクラスタに誤分類されることになる。

【0011】この誤分類を防ぐためには、従来の方式では、プロトタイプデータも含めて、すべてのデータを用いてクラスタリングし直す必要があり、大きな作業負担が強いられる。データを新たに追加した場合にクラスタの修正を行う方法が、特開平5-205058号に開示されているが、これは、新たなデータを追加したこと 既知でなければならず、かつ外部からデータの追加によるクラスタの修正を実行することを知らせる必要があり、追加するデータを自動的に集めたり、クラスタを自動的に修正することはできない。

【0012】本発明は、こうした従来の問題点を解決するものであり、簡単な構成と手順で、クラスタリングにおけるデータの動的変化に対応することができるクラスタリング装置を提供することを目的としている。

【0013】

【課題を解決するための手段】そこで、本発明では、入力データを、クラスタを用いて分類するクラスタリング装置において、クラスタを作成するクラスタ作成装置と、クラスタ作成装置により作成されたクラスタを用いて、入力データのクラスタリングを実行するクラスタリング実行装置と、クラスタリング実行装置のクラスタリング結果を監視して誤分類された入力データを識別するクラスタリング結果監視装置と、誤分類された入力データを蓄積する蓄積手段とを設け、蓄積手段に一定数以上のデータが蓄積された場合に、このデータを基に、クラスタ作成装置が新たなクラスタを作成するように構成している。

【0014】そのため、入力データの動的変化に対応してクラスタを修正し、誤分類を抑えることができる。

【0015】

【発明の実施の形態】以下、本発明の実施の形態について、図面を用いて説明する。なお、本発明はこれら実施の形態に何等限定されるものではなく、その要旨を逸脱しない範囲において種々なる態様で実施し得る。

【0016】(第1の実施形態)第1の実施形態のクラスタリング装置は、図1に示すように、プロトタイプデータを管理するプロトタイプデータDB4と、プロトタイプデータを用いてクラスタ5を作成するクラスタ作成

装置1と、作成されたクラスタ5を用いて入力データ6をクラスタリングするクラスタリング実行装置2と、クラスタリング実行装置2のクラスタリング結果7を監視するクラスタリング結果監視装置3と、クラスタリング結果監視装置3によって誤分類と判断されたデータを管理する誤分類入力データDB8とを備えている。

【0017】この装置では、クラスタ作成装置1が、プロトタイプデータDB4を用いてクラスタ5を生成する。クラスタリング実行装置2は、生成されたクラスタ5を用いて、入力された入力データ6をクラスタリングし、クラスタリング結果7を出力する。クラスタリング結果監視装置3は、出力されたクラスタリング結果7を監視し、入力データ6のクラスタリング結果7に含まれる誤差が、ある一定値以上の値であり、明らかに誤分類であると判断した時、その入力データ6を誤分類入力データDB8に追加し、誤分類入力データDB8に溜まったデータの数をカウントする。この誤分類入力データDB8内のデータがある一定数を超えた時、クラスタ作成装置1に、この誤分類入力データDB8を用いて、クラスタを作成するように指示する。

【0018】各装置の動作をさらに詳しく説明する。まず、クラスタ作成装置1は、クラスタ5が作成されていない時と、クラスタリング結果監視装置3からクラスタの作成を指示された時に動作する。

【0019】クラスタ5が作成されていない時は、プロトタイプデータDB4内のプロトタイプデータ群の個々のデータを初期クラスタと見なし、その中から、最も距離の近いクラスタを探す。この距離は図5の式1によって求める。この時求められた2つのクラスタを統合し新たなクラスタとする。統合されたクラスタは削除され、また新たに作られたクラスタは、統合により削除されたクラスタの値をすべて持つ。同様にまたプロトタイプデータDBから、最も距離の近いクラスタを探し、それらを統合する、という一連の作業を繰り返す。この時、全クラスタ数が1になった場合や、最も距離の近いクラスタ同士の距離が、ある一定値より大きかった場合は、作業を終了する。

【0020】この一連の作業により、プロトタイプデータ4から作成されたクラスタ5を作成する。

【0021】次に、クラスタリング結果監視装置3からクラスタ作成の指示を受けた時は、誤分類入力データDB8を用い、クラスタ5を作成するのと同じ動作で、クラスタを作成する。この時、作成されたクラスタで、クラスタ内に含まれる値の数が一定以上のものを新たなクラスタとしてクラスタ5に加える。最後に、誤分類入力データDBをクリアする。

【0022】次に、クラスタリング実行装置2の動作について説明する。クラスタ作成装置1により作成されたクラスタ5を用いて、入力された入力データ6と、距離の最も近いクラスタを選択する。この距離の計算は、図

5の式1によって求める。この時、選択されたクラスタと、誤差を表す、計算された距離とをクラスタリング結果7として出力する。

【0023】クラスタリング結果監視装置3は、出力されたクラスタリング結果7に含まれる誤差、即ち、計算された距離が、ある一定値以上の値である時、誤分類入力データDB8に入力データ6を追加しその数をカウントし、この誤分類入力データDB8内のデータがある一定数を超えた時、クラスタ作成装置1にこの誤分類入力データDB8を用いて、クラスタを作成するように指示する。

【0024】以上のように、この実施形態のクラスタリング装置では、稼動中にもクラスタの自動作成が可能であり、入力データの動的な変化に対応して自動的にクラスタを作成することができる。そのため、入力データの動的な変化に起因する誤分類の発生が迅速に抑えられる。また、この装置では、クラスタの再作成が、実データのクラスタリングの過程で誤分類データとして自動収集されたデータのみを用いて行われるため、少ない負担でクラスタの修正を実行することができる。

【0025】(第2の実施形態) 第2の実施形態のクラスタリング装置は、自己組織化マップ(以下、SOMと云う)を利用してクラスタを作成する。

【0026】この装置は、図2に示すように、第1の実施形態と同様、プロトタイプデータDB4、クラスタ作成装置1、クラスタリング実行装置2、クラスタリング結果監視装置3及び誤分類入力データDB8から成り、クラスタ作成装置1は、プロトタイプデータを入力するデータ入力手段11と、SOM9を作成するSOM作成手段12と、SOM9を用いてクラスタを生成するクラスタ生成手段13とを備え、また、クラスタリング結果監視装置3は、クラスタリング実行装置2のクラスタリング結果7を監視するクラスタリング結果監視手段31と、誤分類入力データDB8のデータを用いてSOM10を作成するSOM修正手段32とを備えている。

【0027】この装置では、クラスタ作成装置1のデータ入力手段11がプロトタイプデータDB4からデータを入力し、このデータを用いてSOM作成手段12がSOM9を作成し、クラスタ生成手段13が、SOM9を用いてクラスタ5を生成する。クラスタリング実行装置2は、生成されたクラスタ5を用いて入力された入力データ6をクラスタリングし、クラスタリング結果7を出力する。クラスタリング結果監視装置3のクラスタリング結果監視手段31は、出力されたクラスタリング結果7に含まれる誤差が、ある一定値以上の値であり、明らかに誤分類であると判断した時、誤分類入力データDB8に入力データ6を追加し、その数をカウントする。

【0028】誤分類入力データDB8内のデータがある一定数を超えた時、SOM修正手段32は、誤分類入力データDB8のデータを入力として新たなSOM10を作成

し、クラスタ作成手段13にSOM10を用いたクラスタ作成を指示する。これを受けて、クラスタ作成手段13は、SOM10を用いてクラスタを作成し、既に作成されているクラスタ5に追加する。

【0029】次に、各部の動作についてさらに詳しく説明する。まず、SOM作成手段12の動作について説明する。

【0030】SOMは、図4に示すように、2次元上に配置されたニューロン402から形成され、各ニューロンは、参照ベクトル403と呼ばれる入力と同じ次元のベクトルを持つ。

【0031】SOM作成手段12は、図3のフローチャートに示す手順でSOMを作成する。

ステップA1：学習回数Tを0にセットし、
ステップA2：図4のように2次元上に配置したニューロンを作成し、各ニューロンに対し、入力と同じ次元の参照ベクトルを乱数で与える。

【0032】ステップA3：データ入力手段11がプロトタイプデータDB4からランダムでデータを一つ取り出す。

【0033】ステップA4：このデータに対して、図5の式(2)を満たす参照ベクトルを持つニューロンCを決定する。

【0034】ステップA5：ニューロンCの近傍に位置するニューロンの参照ベクトルを、図5の式(3)に従って更新する。

【0035】ステップA6：学習回数Tが規定した回数に達した場合には、

ステップA8：終了する。

【0036】ステップA6において、学習回数Tが規定回数に達していない場合には、

ステップA7：学習回数Tの値を一つ増やし、ステップA2に戻る。

【0037】次に、クラスタ生成手段13は、クラスタ5が作成されていない時と、SOM修正手段32からクラスタの作成の指示を受けた時に動作する。

【0038】まず、クラスタ5が作成されていない時、SOM9を用いてクラスタ5を作成する。SOM9の各ニューロンに対し、図5の式(4)を満たす参照ベクトルを持つニューロンを選択し、選択されたニューロンを初期クラスタと見なす。その中から、最も距離の近いクラスタを探す。この距離は図5の式(1)によって求める。この時求められた2つのクラスタを統合し新たなクラスタとする。統合されたクラスタは削除し、また新たに作られたクラスタは、統合により削除されたクラスタの値をすべて持つ。同様にまた、最も距離の近いクラスタを探し、それらを統合する、という一連の作業を繰り返す。この時、全クラスタ数が1になった場合や、最も距離の近いクラスタ同士の距離が、ある一定値より大きかった場合は、作業を終了する。

【0039】また、SOM修正手段32からクラスタの作成の伝達を受けた時も同様に、SOM10を用いてクラスタを作成し、クラスタ5に追加をする。

【0040】クラスタリング実行装置2は、第1の実施形態と同様、クラスタ5を用いて入力データ6をクラスタリングし、クラスタリング結果7を出力する。クラスタリング結果監視手段31は、出力されたクラスタリング結果7に含まれる誤差が、ある一定値以上の値であり、明らかに誤分類であると判断した時、誤分類入力データDB8に入力データ6を追加し、その数をカウントする。

【0041】この誤分類入力データDB8内のデータがある一定数を超えた時、SOM修正手段32は、誤分類入力データDB8のデータを入力として、図3のフローチャートに従って、マップの大きさがSOM9の縦または横のニューロンの数と等しい、小さいSOM10を作成する。そして、誤分類入力データDB8をクリアし、クラスタ作成手段13にクラスタの作成を指示する。クラスタ作成手段13は、前述するように、SOM10を用いてクラスタを作成し、作成済みのクラスタ5に追加する。

【0042】以上のように、この実施形態のクラスタリング装置では、SOMを用いてクラスタリングを行っているため、既存のSOMをそのまま適用することができ、さらに新たにクラスタを作成する際に非常に小さいSOMを用いるので処理速度も高く、その実用的效果は大きい。また、この新たなクラスタの作成には、実データのクラスタリングの過程で誤分類データとして自動収集されたものが使用されるため、この新たなクラスタの作成により、入力データの動的な変化に対応することができる。

【0043】

【発明の効果】以上の説明から明らかなように、本発明のクラスタリング装置は、入力データの動的な変化に対応して、新たなクラスタを速やかに作成することができ、入力データの動的な変化に起因する誤分類の発生を抑えることが可能である。

【0044】また、この新たなクラスタの作成は、クラスタリングを実行したときに、誤分類データとして自動収集されたデータだけを用いて行われるため、その作成負担は少なくて済む。

【0045】また、誤分類されたデータからクラスタを直接作成する手段を持つ装置では、装置稼動中にもクラスタを自動で作成することが可能であり、入力データの動的な変化に素早く対応できるという有利な効果が得られる。

【0046】また、SOMを用いてクラスタリングする装置では、既存のSOMをそのまま適用することができ、また、新たにクラスタを作成する際には非常に小さいSOMを用いるので処理速度も高いという有効な効果が得られる。

10

【0047】このことにより、本発明は、入力データが時間的に変化するものをクラスタリングする装置に適用して効果を発揮することができ、例えば、時間的に変化する生徒の学習結果を入力データとして生徒を分類する学習システムのクラスタリング装置や、インターネットのホームページにアクセスする視聴者の嗜好性を調査するクラスタリング装置などに用いた場合に、極めて有効である。

【図面の簡単な説明】

10

【図1】本発明の第1の実施形態におけるクラスタリング装置の構成を表すブロック図、

【図2】本発明の第2の実施形態におけるクラスタリング装置の構成を示すブロック図、

【図3】第2の実施形態においてSOM作成の手順を示すフローチャート、

【図4】SOMを視覚的に示す図、

【図5】数式を示す図、

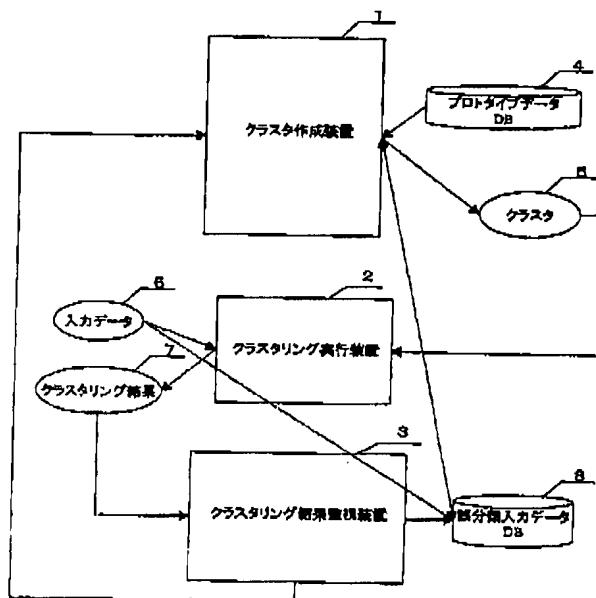
【図6】従来のクラスタリング装置の一例を示す図である。

20

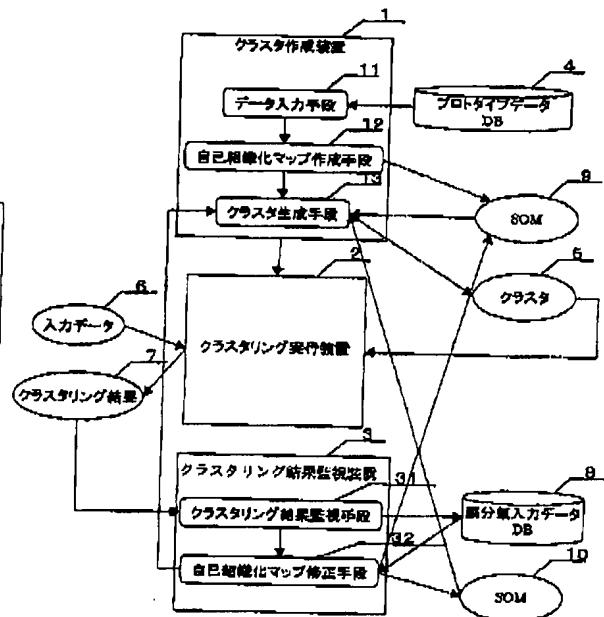
【符号の説明】

- 1 クラスタ作成装置
- 2 クラスタリング実行装置
- 3 クラスタリング結果監視装置
- 4 プロトタイプデータDB
- 5 クラスタ
- 6 入力データ
- 7 クラスタリング結果
- 8 誤分類入力データDB
- 9、10 SOM
- 30 11 データ入力手段
- 12 SOM作成手段
- 13 クラスタ生成手段
- 31 クラスタリング結果監視手段
- 32 SOM修正手段
- 100 プロトタイプデータ群
- 102 クラスタA
- 103 クラスタB
- 104 距離
- 105 クラスタC
- 40 200 プロトタイプデータから作成されたクラスタ結果
- 201 クラスタY
- 202 クラスタZ
- 300 クラスタを用いたクラスタリング装置
- 301 クラスタY
- 302 クラスタZ
- 303 入力X
- 304 入力Xのポイント
- 401 SOM
- 402 ニューロン
- 50 403 参照ベクトル

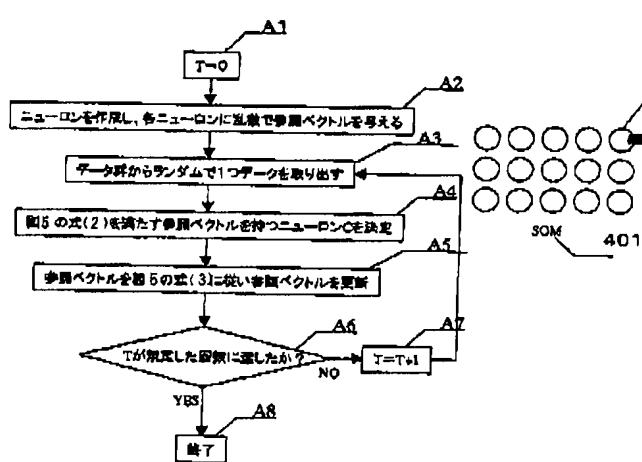
【図1】



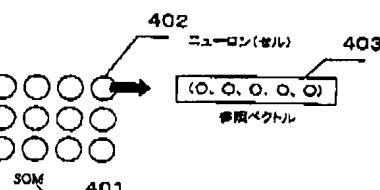
【図2】



【図3】



【図4】



(7)

特開2001-283184

【図5】

N次元のデータX,Yの距離D $D = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$ (1)

 $C = \arg \min_{i \in N_c} \|x - m_i\|$ (2)
 i:ニューロン番号
 X:入力データ
 m_i :ニューロンiの参照ベクトル
 $\|\cdot\|$:式(1)によって求められる距離

$$m_i = \begin{cases} m_i + h_\alpha(t)[x(t) - w_i(t)] & (i \in N_c) \\ m_i & (i \notin N_c) \end{cases}$$
 (3)

$$h_\alpha = \alpha(t) \cdot \exp \left[-\frac{\|t_c - t_j\|^2}{2\sigma^2(t)} \right]$$

N_c :ニューロンCの近傍に位置するニューロン
 i:ニューロン番号
 α :学習係数:0 < α < 1 (時間とともに減少)
 σ :近傍幅 (時間とともに減少)
 $\|\cdot\|$:式(1)によって求められる距離
 t_c :2次元マップ上のニューロンCの位置ベクトル
 t_j :2次元マップ上のニューロンiの位置ベクトル

 $d_i < \min_{j \in N_i} d_j$ (4)
 $d_i = \frac{1}{|N_i|} \sum_{j \in N_i} \|w_i - m_j\|$

N_i :ニューロンiの4近傍中に存在するニューロンの集合
 m_i :ニューロンiの参照ベクトル
 w_i :ニューロンiの参照ベクトル
 $\|\cdot\|$:式(1)によって求められる距離

【図6】

